基于Patch域对抗训练的语音增强

王鸿韬，陆志华，叶庆卫，章联军
（宁波大学信息科学与工程学院，浙江 宁波 315211）

摘　要：在基于深度学习的语音增强方法中，往往会遇到训练数据和测试数据分布不匹配的问题，这种不匹配包括两个数据中说话人、说话内容、噪声类型及信噪比的不匹配。严重的数据不匹配问题会导致语音增强的性能大幅下降，针对这种情况提出了一种基于Patch域对抗训练的语音增强方法。该方法在先前域对抗训练的语音增强方法基础上，通过域判别器的隐式建模，能使整段语音被划分为多个独立Patch再进行判别，实现了对训练数据的适应性学习，从而减小训练数据和测试数据之间的分布差异，提高了模型在测试数据上的增强能力。实验结果表明，该方法在不同程度的数据不匹配问题下较先前方法都表现出优异的性能，且作为对抗训练也保持了良好的稳定性。

关键词：语音增强；域对抗训练；领域自适应
中图分类号：TN915
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2024225

Patch-based domain adversarial training for speech enhancement

WANG Hongtao, LU Zhihua, YE Qingwei, ZHANG Lianjun
Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo 315211, China

Abstract: In deep learning-based speech enhancement methods, mismatched distributions between training data and test data are often encountered. These mismatches can include differences in speakers, speech content, noise types, and signal-to-noise ratios between the datasets. Severe data mismatches can significantly degrade the performance of speech enhancement. To address this issue, a speech enhancement method based on Patch domain adversarial training was proposed. Building on previous domain adversarial training methods for speech enhancement, implicit modeling of a domain discriminator was employed, allowing the entire speech signal to be divided into multiple independent patches for discrimination. Adaptive learning of the training data was enabled, thereby reducing distribution differences between the training and test data and improving the model's enhancement capabilities on test data. Experimental results show that this method exhibits superior performance compared to previous methods under various degrees of data mismatch and maintains good stability as an adversarial training approach.

Key words: speech enhancement, domain adversarial training, domain adaptation

收稿日期：2024−07−19；修回日期：2024−09−09

0　引言

语音增强（speech enhancement，SE）是一种减少语音信号中的噪声、提高语音质量的技术。SE在与语音相关应用中被广泛使用，例如自动语音识别、声纹识别、助听器设计和语音编码等。传统的语音增强方法包括谱减法[1]、维纳滤波器[2]和非负矩阵分解（non-negative matrix factorization，NMF）[3]等，这些方法以信号处理的方式，基于一定的假设对语音信号进行建模，但现实声学场景中的语音信号更为复杂，这些方法不能获得很好的效果。近年来，随着计算机硬件的迅速发展，深度学习在语音增强领域表现出令人满意的成果。在有监督的语音增强方法中通常采用掩码（masking）、映射（mapping）和聚类（clustering），这些方法都是基于数据驱动的深度神经网络（deep neural network，DNN）[4]，对语音特征进行深度建模，以适应复杂的声学环境。然而，在实际应用中，基于深度学习的语音增强方法常常会面临训练和测试环境不匹配的问题。这种不匹配问题会严重影响算法性能，因为训练数据和测试数据分布并不一致，这更符合实际情况。另外，训练数据的采集也相当困难，涉及专业设备和实际场景，难以获取涵盖各种声学场景的大量干净-带噪并行的成对语音数据。因此，如何利用有限的训练数据提高算法在测试数据上的泛化能力成了一个较为棘手的问题，领域自适应（domain adaptation，DA）[5]是解决这一问题的有效手段。

DA在计算机视觉领域取得了不错的效果，其主要思想是将无标签的测试数据（目标域数据）加入有标签的训练数据（源域数据）中一起训练，将模型从源域迁移到目标域。DA的一类做法是通过域对抗训练（domain adversarial training，DAT）[6]提取源域和目标域的域不变特征，从而提高模型在新领域的鲁棒性，减少两个域之间的域漂移（domain shift）。在语音增强领域，源域数据为干净-带噪并行的成对语音数据，而目标域数据仅包含带噪语音。

Hou等[7]和Liao等[8]都将DAT引入语音增强领域，本文将其方法都称为DAT-SE（domain adversarial training for speech enhancement），但实现方式略不同。Hou等[7]对训练数据进行复杂处理，分为静态特征、一阶特征和二阶特征，语音增强方式选择了掩码，判别方式是判别语音来自哪个域；Liao等[8]则是对训练数据进行大量切片，将完整的语音分为多个小块进行判别，对噪声类型进行判别，语音增强方式选择了映射。由此可见，无论是哪种域对抗训练方法，都对训练数据进行了复杂处理，这和有监督的语音增强方法背道而驰。对于多类特征，这样的处理方式是烦琐的，还存在特征冗余问题；对于切片特征，破坏了语音自身存在的长序列特点，只关注了块内特征，而忽视了块间特征，即只关注了局部特征而忽略了全局特征，在一些双路径建模的语音增强方法中特别强调了这一点，如GALR[9]、SepFormer[10]等。

综上，在缓解训练数据和测试数据不匹配问题时，DAT-SE对训练数据预处理步骤复杂且会出现特征冗余和破坏全局特征的情况，本文提出一种基于Patch域对抗训练的语音增强方法，称为PDAT-SE（patch-based domain adaptation for speech enhancement）。PDAT-SE的基本思想是将整段语音映射到高维的特征空间中，将长序列语音特征分为多个Patch，让判别器对每个Patch进行得分估计，从而获取整段语音的综合得分。该方式既提高了模型增强能力，也提高了模型判别能力，使模型能够提取源域和目标域之间更多的域不变特征，从而减少源域和目标域之间的分布差异。

本文研究做出了以下3个方面的贡献。首先，用域判别器隐式建模的方式将整段语音划分为多个

独立Patch进行判别，为判别语音特征提供了新思路。其次，虽然对抗模型本身难以训练，但PDAT-SE延续了DAT-SE的稳定性，且在性能方面的表现也更加优异。另外，PDAT-SE简化了数据的预处理步骤，使得该方法和有监督的语音增强方法处理数据方式一致，为对抗训练在语音增强中的应用提供了一个更为通用的方法。

1 相关工作

1.1 领域自适应

给定一个有标签源域$\mathcal{D}_s=\{x_i,y_i\}_{i=1}^{N}$表示干净-带噪并行的成对语音集合，其中$x_i$和$y_i$分别表示源域的带噪语音及其对应的干净语音。给定无标签目标域$\mathcal{D}_t=\{x_j\}_{j=N+1}^{N+M}$无并行的语音集合，其中$x_j$表示目标域的带噪语音。

领域自适应的常用方法是通过学习特征映射，将源域和目标域样本投影到某一个特征空间上，使得在这个空间下两者分布接近，可以表示为：

$$p(M_s(X_s))\approx p(M_t(X_t)) \tag{1}$$

其中，X_s和X_t分别表示源域样本和目标域样本的集合，即$x_i\in X_s$，$x_j\in X_t$，$M_s(\cdot)$和$M_t(\cdot)$分别表示源域和目标域的特征映射，$p(\cdot)$表示概率分布。模型自适应后，在$M_s(X_s)$上训练好的模型可以直接用于目标域。

1.2 语音增强

在语音增强领域中，语音信号可以表示为：

$$x(t)=y(t)+n(t) \tag{2}$$

其中，$x(t)$表示带噪语音，$y(t)$表示干净语音，$n(t)$表示噪声。对式（1）进行短时傅里叶变换（short-time Fourier transform，STFT）：

$$x(k,f)=y(k,f)+n(k,f) \tag{3}$$

其中，$k\in[1,K]$表示时间帧索引，$f\in[1,F]$表示频率索引，K和F分别表示时间帧和频率的总数。$x(k,f)$、$y(k,f)$、$n(n,f)$分别对应时域$x(t)$、$y(t)$、$n(t)$短时傅里叶变换后的频谱。本文均用频谱作为SE模型的输入和输出，为了下文表达的简洁性，省略了时间帧索引和频率索引。

对于SE框架来说，可以理解为对SE模型$h(\cdot;\theta_{\text{se}})$寻找最优参数集合$\theta_{\text{se}}^{*}$，使得增强损失$\mathcal{L}_{\text{se}}$最小，可以表示为：

$$\min_{\theta_{\text{se}}}\mathcal{L}_{\text{se}}=\min_{\theta_{\text{se}}}\frac{1}{N}\sum_{i=1}^{N}D(h(x_i;\theta_{\text{se}}),y_i) \tag{4}$$

其中，$D(\cdot,\cdot)$是距离度量，可以有多种选择，例如平均误差（mean absolute error，MAE）、均方误差（mean squared error，MSE）等，其目的都是测算干净语音y和经SE模型增强后语音$\hat{x}$之间的相似度。

SE框架由一个编码器（encoder）和一个解码器（decoder）组成，编码器提取带噪语音特征，解码器通过特征再恢复为干净语音。式（4）可以改写成带参数形式：

$$\min_{\theta_{\text{enc}},\theta_{\text{dec}}}\mathcal{L}_{\text{se}}=\min_{\theta_{\text{enc}},\theta_{\text{dec}}}\frac{1}{N}\sum_{i=1}^{N}D(G_{\text{dec}}(G_{\text{enc}}(x_i;\theta_{\text{enc}});\theta_{\text{dec}}),y_i) \tag{5}$$

其中，G_{enc}和G_{dec}分别表示编码器和解码器，θ_{se}由编码器的参数θ_{enc}和解码器的参数θ_{dec}组成，即$\theta_{\text{se}}=\{\theta_{\text{enc}},\theta_{\text{dec}}\}$。

2 基于Patch域对抗训练的语音增强

在原始的生成对抗网络（generative adversarial network，GAN）[11]中，有一些研究[12]已经表明要对高分辨率、高清细节的图片进行判别并不合适，因为在这些GAN中的判别器都是对整张图片进行判别。鉴于此，也有许多将GAN应用到语音增强领域的工作，如CMGAN[13]、SCP-GAN[14]等，但这些工作也都是对整个语音进行判断。因此，本文提出可以将语音划分为多个独立Patch再进行判别的PDAT-SE，基于Patch域对抗训练的语音增强（PDAT-SE）整体框架如图1所示，

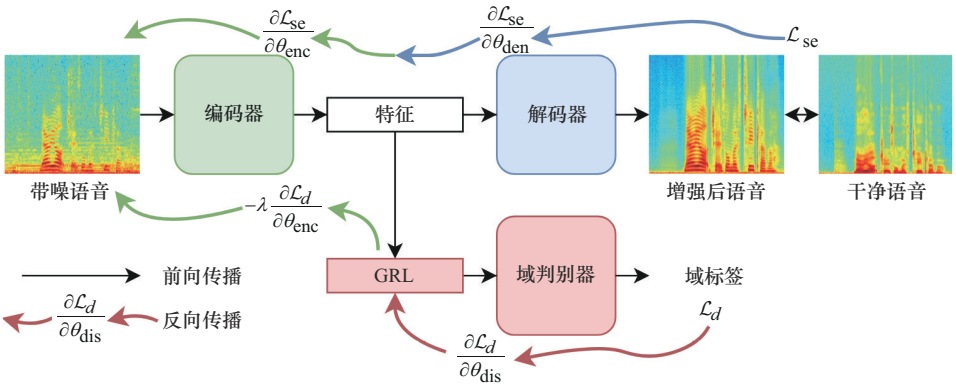

图 1　基于 Patch 域对抗训练的语音增强（PDAT-SE）整体框架

主要包括 3 部分：一是由编码器和解码器组成的 SE 通用框架，其任务和 SE 一致，用来恢复干净语音；二是域判别器（domain discriminator），用来判别特征来自哪个域；三是梯度反转层（gradient reversal layer，GRL），GRL 能够使梯度以负的形式反向传播。

2.1　基于 Patch 的域判别器

在域判别器中，用指示函数 $\mathbf{1}_{\mathcal{D}_s,\mathcal{D}_t}$ 作为样本来源的标签，如果样本来自源域，用 0 表示，来自目标域则用 1 表示。因此指示函数 $\mathbf{1}_{\mathcal{D}_s,\mathcal{D}_t}$ 可以写成：

$$\mathbf{1}_{\mathcal{D}_s,\mathcal{D}_t}(x)=\begin{cases}0, & x\in\mathcal{D}_s\\ 1, & x\in\mathcal{D}_t\end{cases} \tag{6}$$

由此，域判别器的交叉熵损失可以定义为：

$$\mathcal{L}_d=-\frac{1}{N+M}\sum_{i=1}^{N+M}[\mathbf{1}_{\mathcal{D}_s,\mathcal{D}_t}(x_i)\log P(x_i)+(1-\mathbf{1}_{\mathcal{D}_s,\mathcal{D}_t}(x_i))\log P(x_i)] \tag{7}$$

将式（7）改写成带参数形式：

$$\mathcal{L}_d=-\frac{1}{N+M}\sum_{i=1}^{N+M}\mathcal{H}(G_{dis}(G_{enc}(x_i;\theta_{enc});\theta_{dis}),\mathbf{1}_{\mathcal{D}_s,\mathcal{D}_t}(x_i)) \tag{8}$$

其中，$\mathcal{H}(a,\hat{a})$ 表示交叉熵，即 $\mathcal{H}(a,\hat{a})=\hat{a}\log P(a)+(1-\hat{a})\log P(a)$，$G_{dis}$ 和 θ_{dis} 分别表示域判别器及其参数。

PDAT-SE 对域判别器的设计灵感来源于图像领域中的 PatchGAN[15]。PatchGAN 设计了一种全卷积判别器，能够对一张图片只在多个 Patch 范围内对其结构进行惩罚。例如，一张图片（大小为 $H\times W$）经过判别器后，输出一个得分图（score map，大小为 $h\times w$），那么对于得分图上的每个元素，实际上可以根据卷积过程追溯到对应原始图片中的各个 Patch（大小为 $R\times R$），因此 Patch 也可以认为是感受野。由此，PatchGAN 的判别器能对图片中的多个区域以 Pacth 的形式进行评价，并对所有结果取均值，通过该方式提升判别器对图片的判别能力。注意到 Patch 的大小都是正方形的，这对于完整频谱（大小为 $F\times K$）来说，一个 $R\times R$ 的 Patch 无法表示语音的任何物理含义。另外，输入 PatchGAN 的判别器是一个原始图像，而在 PDAT-SE 的域判别器中，它是对特征空间中的高维语音特征进行判别。因此，PDAT-SE 对增强模块和域判别器都进行了调整，规避了上述问题。

在 PDAT-SE 的增强模块中，延续了一些 SOTA（state of the art）的语音增强方法的处理方式，例如，DBT-Net[16]、MP-SENet[17]等方法中，大多采用频谱图作为网络的输入，并且在涉及上下采样时，保持了时间帧 K 的不变性，而只对频率 F 进行上下采样，该方式能保证语音信号的连续性，不丢失长序列信息。PDAT-SE 的域判别器

在接收到时间帧K不变但高维频率F_{h}的频谱特征图时，将其频率压缩至一个较低维度F_l，对时间帧以L帧、步长为$L/2$对其进行二次压缩，可以表示为：

$$W_{\mathrm{output}}=C_2\left(C_1\left(W_{\mathrm{input}},\theta_{c_1}\right),\theta_{c_2}\right) \tag{9}$$

其中，$W_{\mathrm{input}}\in\mathbb{R}^{K\times F_h}$表示输入第一次压缩网络$C_1(\cdot;\theta_{c_1})$前的频谱特征图，$C_2(\cdot;\theta_{c_2})$表示第二次压缩网络，$\theta_{c_1}$、$\theta_{c_2}$各表示其压缩网络的参数，且$\theta_{c_1},\theta_{c_2}\in\theta_{\mathrm{dis}}$。压缩网络可以由某个模块操作完成，不局限于卷积、池化等。从而，可以得到一个K'长的1维序列$W_{\mathrm{output}}\in\mathbb{R}^{K'\times 1}$，该序列每个元素对应着$L\times F$大小的频谱图，也对应着时域语音的一段波形信号。最后对该1维序列取均值，得到该段语音的综合判别结果。

2.2 联合对抗训练

现对PDAT-SE进行联合训练，同时实现语音增强和域判别的任务。对于语音增强任务来说，通过寻找最优参数θ_{enc}、θ_{dec}使得$\mathcal{L}_{\mathrm{se}}$最小。对于域判别任务来说，PDAT-SE一方面希望寻找最优参数θ_{enc}以最大化域判别损失$\mathcal{L}_d$，从而使域判别器分不清样本来自哪个域，即说明语音被映射到一个源域分布和目标域分布接近的特征空间，这样编码器提取的都是共同特征；另一方面，PDAT-SE同时希望寻找最优参数θ_{dec}以最小化域判别损失$\mathcal{L}_d$，即希望域判别器能够判别样本来自哪个域。这种对抗形式的优化可以通过GRL来实现，梯度在前向传播时是其本身，在反向传播时是其本身的$-\lambda$倍。因此，PDAT-SE联合训练的损失可以写成：

$$\mathcal{L}=\mathcal{L}_{\mathrm{se}}-\lambda\mathcal{L}_d \tag{10}$$

结合式（5）和式（8），将式（10）改写成带参数形式：

$$\mathcal{L}\left(\theta_{\mathrm{enc}},\theta_{\mathrm{dec}},\theta_{\mathrm{dis}}\right)=\mathcal{L}_{\mathrm{se}}\left(\theta_{\mathrm{enc}},\theta_{\mathrm{dec}}\right)-\lambda\mathcal{L}_d\left(\theta_{\mathrm{enc}},\theta_{\mathrm{dis}}\right) \tag{11}$$

那么式（11）中各个参数的求解可以表示为：

$$\begin{aligned}&(\hat{\theta}_{\mathrm{enc}},\hat{\theta}_{\mathrm{dec}})=\arg\min_{\theta_{\mathrm{enc}},\theta_{\mathrm{dec}}}\mathcal{L}(\theta_{\mathrm{enc}},\theta_{\mathrm{dec}},\hat{\theta}_{\mathrm{dis}})\\&\hat{\theta}_{\mathrm{dis}}=\arg\max_{\theta_{\mathrm{dis}}}\mathcal{L}(\hat{\theta}_{\mathrm{enc}},\hat{\theta}_{\mathrm{dec}},\theta_{\mathrm{dis}})\end{aligned} \tag{12}$$

若用梯度下降法对式（12）中的参数进行更新，更新规则如下。

$$\begin{aligned}\theta_{\mathrm{enc}}&\leftarrow\theta_{\mathrm{enc}}-\mu\left(\frac{\partial\mathcal{L}_{\mathrm{se}}}{\partial\theta_{\mathrm{enc}}}-\lambda\frac{\partial\mathcal{L}_d}{\theta_{\mathrm{enc}}}\right)\\\theta_{\mathrm{dec}}&\leftarrow\theta_{\mathrm{dec}}-\mu\frac{\partial\mathcal{L}_{\mathrm{se}}}{\theta_{\mathrm{dec}}}\\\theta_{\mathrm{dis}}&\leftarrow\theta_{\mathrm{dis}}-\mu\frac{\partial\mathcal{L}_d}{\theta_{\mathrm{dis}}}\end{aligned} \tag{13}$$

其中，μ为学习率，λ为GRL的自适应系数。

在测试阶段，需要丢弃域判别器，只需要实现语音增强的任务。

3 实验结果与分析

3.1 实验数据集

本文实验使用的干净语音均来自VCTK数据集[18]，噪声来自Demand数据库[19]和人为噪声。为了更符合实际情况，本文根据数据不匹配问题的严重程度制作了两个域漂移程度不同的数据集，即域漂移较小的数据集和域漂移较大的数据集，分别记作LDSD（low domain shift dataset）和HDSD（high domain shift dataset）。

在LDSD中，总共考虑了40种不同条件：10种噪声（8种来自Demand和2种人为噪声），每种噪声有4种信噪比，即{0, 5, 10, 15}dB。说话人包括14名男性和14名女性，经混合后总共有11 572个成对语音。在该数据集中以噪声类型来区分源域数据和目标域数据，在10种噪声中选取5种作为源域数据，共5 780个成对语音，剩下5种作为目标域数据，共5 792个成对语音，为了验证PDAT-SE在减少域间漂移的有效性，对于测试集数据，随机从目标域中选取450个成对语音，所以实际训练用到的目标域数据为5 342个。在

该数据集中，源域数据和目标域数据实现了信噪比和说话人一致，但说话内容和噪声类型不一致，两者有一定差异。

在HDSD中，为了扩大域之间的差异，该数据集从VCTK的110个说话人中随机选取6名男性和6名女性的干净语音，从Demand中随机选取5种噪声，以17.5 dB的信噪比进行混合作为源域数据，共4 859个成对语音。对于目标域数据，以相同方式选取与源域不重复的说话人和噪声类型，并以4种信噪比，即{0, 5, 10, 15}dB进行混合，共5 146个成对语音。对于测试集数据，随机从目标域中选取287个成对语音，使源域数据和目标域数据数量一致。因此在该数据集中，源域数据和目标域数据在所有情况包括信噪比、说话人、说话内容和噪声类型上都不一致，两者有较大差异。

数据集LDSD和HDSD的区别见表1，一致用T表示，不一致用F表示。

表 1　数据集LDSD和HDSD的区别

数据集	信噪比	说话人	说话内容	噪声类型
LDSD	T	T	F	F
HDSD	F	F	F	F

3.2　实验设置

本文使用BLSTM[20]模型作为SE网络的构建模块。编码器和解码器各包含一个具有512个节点的双向LSTM层，解码器在输出端包含一个具有257个线性节点的全连接层，用于频谱图估计。本文所提域判别器是一个具有1 024个节点的单向LSTM和两个池化层，sigmoid层用于Patch的得分估计。

本文对语音的采样频率为16 kHz，使用512点STFT提取时频（T-F）特征，其汉明窗大小为32 ms，帧移为16 ms，从而得到由257点STFT对数功率谱（logarithmic power spectrum，LPS）组成的特征向量。对于DAT-SE，将每个训练语音分成32帧的多个片段，因此编码器和解码器的输入输出都是一个257×32的矩阵。最后，通过傅里叶反变换和叠加法将多个解码器输出串接并合成回波形信号；对于PDAT-SE，简化DAT-SE对训练语音的处理方式，直接将特征向量（频谱）输入编解码器，再通过傅里叶反变换到时域波形信号即可。另外，两者都是用噪声信号的相位进行傅里叶反变换。

本文使用Adam优化器进行训练，共训练300个epoch，SE模块和域判别器的学习率分别为0.000 1和0.000 5。设置批量大小为16，自适应系数$\lambda=0.05$。

3.3　评价指标

本文共使用以下6种客观指标来评估算法性能，分别是：语音质量的感知评估（PESQ）[21]，得分范围为−0.5到4.5；短时客观可理解度（STOI）[22]，得分范围为0到1，一般以“%”的形式表现；主观平均意见评分（MOS）[23]，包括信号失真的CSIG评分、噪声失真的CBAK评分和综合质量的COVL评分，所有MOS的得分范围都是1到5；分段信噪比（SSNR）[23]，取值范围为−10到35。

3.4　实验结果

在LDSD测试集上的算法性能见表2，表2中Noisy代表测试集的带噪语音，Source表示只用源域数据进行训练，Target表示只用带标签的目标域数据进行训练，这表示算法的性能上限，加粗表示除用目标域数据训练外的最好数值。

表 2　在LDSD测试集上的算法性能

方法	PESQ	STOI	CSIG	CBAK	COVL	SSNR
Noisy	1.30	81.82	2.19	1.88	1.69	-0.71
Source	1.95	82.69	3.25	2.56	2.57	3.56
DAT-SE	2.02	82.96	3.44	2.62	2.71	3.78
PDAT-SE	**2.09**	**84.07**	**3.49**	**2.67**	**2.77**	**4.06**
Target	2.20	85.62	3.63	2.80	2.90	4.91

为了更明显地体现DAT-SE和PDAT-SE的性能差异，本文定义了减小域漂移百分比（reduce domain shift percentage，RDSP），用E_{RDSP}表示：

$$E_{\mathrm{RDSP}}=\frac{g(\mathrm{Method})-g(\mathrm{Source})}{g(\mathrm{Target})-g(\mathrm{Source})} \tag{14}$$

其中，Method表示DAT-SE或PDAT-SE，$g(\cdot)$表示某方法在测试集上同一评价指标下取得的具体数值。$g(\mathrm{Target})-g(\mathrm{Source})$就是目标域和源域之间产生的域漂移。

根据表2和式（14）画出RDSP图，在LDSD测试集上两种方法的RDSP如图2所示。由表2和图2可以表明，PDAT-SE在各项语音相关指标上都远超DAT-SE，尤其在STOI上，PDAT-SE相比于DAT-SE提升了1.11%，RDSP能减少将近40%。

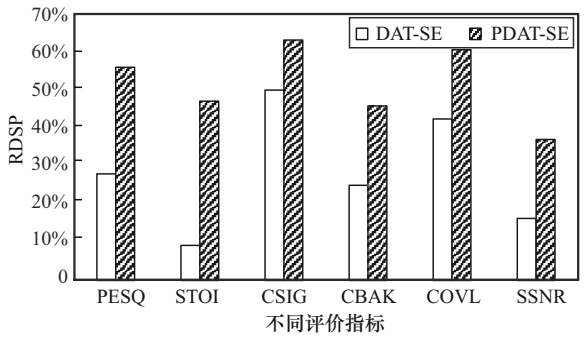

图2　在LDSD测试集上两种方法的RDSP

GRL的自适应系数λ对平衡增强损失$\mathcal{L}_{\mathrm{se}}$和域判别损失$\mathcal{L}_d$起到了关键作用，因此本文另外测试了$\lambda$为0.5和0.005时PDAT-SE的性能，在LDSD测试集上不同λ下的PDAT-SE性能见表3。当$\lambda=0$时，即相当于只用源域数据进行训练。一般情况下，DAT-SE在$\lambda<0.1$时能起作用，在$\lambda=0.05$时性能最佳，由表3可以发现PDAT-SE也是如此。

表3　在LDSD测试集上不同λ下的PDAT-SE性能

λ	PESQ	STOI	CSIG	CBAK	COVL	SSNR
0	1.95	82.69	3.25	2.56	2.57	3.56
0.5	1.93	82.68	3.37	2.53	2.62	3.17
0.05	**2.09**	**84.07**	**3.49**	**2.67**	**2.77**	**4.06**
0.005	2.03	83.37	3.40	2.63	2.69	3.97

在HDSD测试集上不同信噪比下的算法性能见表4，加粗表示DAT-SE和PDAT-SE中数值较高的值。在HDSD中，由于源域语音和目标域语音的信噪比不同，在测试时按目标域语音的信噪比分别测试。由表4可以发现，在低信噪比时，PDAT-SE没有发挥很大作用，但随着信噪比的增加，PDAT-SE与DAT-SE的差异越来越大，优势越来越明显。换句话讲，在低信噪比时，对一段语音划分为多个子块和划分为多个Patch的方式性能接近，这也说明对于低信噪比的语音信号，长序列信息被噪声破坏得较为严重，无论是分块信号输入网络还是整段信号输入网络，网络都无法捕获其长序列信息。而在高信噪比时，PDAT-SE能够捕获远距离信息的优势就能体现出来。

综上所述，无论哪种情况，仅用源域训练的模型也具有一定的泛化能力，PDAT-SE能够比DAT-SE更好地缓解域漂移，减小源域和目标域

表4　在HDSD测试集上不同信噪比下的算法性能

方法	SNR=0			SNR=5 dB			SNR=10 dB			SNR=15 dB		
	PESQ	STOI	SSNR	PESQ	STOI	SSNR	PESQ	STOI	SSNR	PESQ	STOI	SSNR
Noisy	1.11	67.03	−4.28	1.21	69.72	−2.11	1.48	74.05	1.25	1.72	76.81	4.62
Source	1.28	66.18	−2.46	1.60	69.79	1.45	2.12	74.73	4.85	2.54	77.36	7.02
DAT-SE	**1.52**	62.92	**0.33**	**1.84**	68.17	2.31	2.23	73.41	4.57	2.59	76.52	6.32
PDAT-SE	1.42	**66.33**	−0.39	1.77	**70.60**	**2.57**	**2.30**	**75.85**	**4.95**	**2.68**	**77.86**	**6.78**
Target	1.88	71.52	2.54	2.25	72.71	3.85	2.68	76.04	5.44	2.96	78.01	6.63

之间的分布差异，尤其在信噪比也不匹配的情况下，PDAT-SE在高信噪比下的泛化能力更强。

4　结束语

本文提出了基于Patch域对抗训练的语音增强（PDAT-SE）方法，该方法延续了基于域对抗训练的语音增强（DAT-SE）方法，对域判别器进行了改进，使其能够隐式地将整段语音以Patch形式进行判别，从而提高了模型的增强能力和判别能力，减少了源域和目标域之间的分布差异。实验结果表明，该方法虽然是对抗训练，但也具有较好的稳定性，在域偏移较小和较大的情况下都优于原来方法。另外，该方法与有监督的语音增强工作的处理数据方式一致，为对抗训练在语音增强中的应用提供了一个通用方法。

参考文献：

[1] BOLL S. Suppression of acoustic noise in speech using spectral subtraction[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1979, 27(2): 113-120.

[2] CHEN J D, BENESTY J, HUANG Y T, et al. New insights into the noise reduction Wiener filter[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2006, 14(4): 1218-1234.

[3] WILSON K W, RAJ B, SMARAGDIS P, et al. Speech denoising using nonnegative matrix factorization with priors[C]//Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2008: 4029-4032.

[4] OCHIENG P. Deep neural network techniques for monaural speech enhancement and separation: state of the art analysis[J]. Artificial Intelligence Review, 2023, 56(3): 3651-3703.

[5] WANG M, DENG W H. Deep visual domain adaptation: a survey[J]. Neurocomputing, 2018, 312: 135-153.

[6] GANIN Y, USTINOVA E, AJAKAN H, et al. Domain-adversarial training of neural networks[J]. Journal of machine learning research, 2016, 17(59): 1-35.

[7] HOU N N, XU C L, CHNG E S, et al. Domain adversarial training for speech enhancement[C]//Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Piscataway: IEEE Press, 2019: 667-672.

[8] LIAO C F, TSAO Y, LEE H Y, et al. Noise adaptive speech enhancement using domain adversarial training[J]. arXiv preprint arXiv: 1807. 07501, 2018.

[9] LAM M W Y, WANG J, SU D, et al. Effective low-cost time-domain audio separation using globally attentive locally recurrent networks[C]//Proceedings of the 2021 IEEE Spoken Language Technology Workshop(SLT). Piscataway: IEEE Press, 2021: 801-808.

[10] SUBAKAN C, RAVANELLI M, CORNELL S, et al. Exploring self-attention mechanisms for speech separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023(31): 2169-2180.

[11] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J]. Advances in Neural Information Processing Systems, 2014, 27.

[12] LARSEN A B L, SØNDERBY S K, LAROCHELLE H, et al. Autoencoding beyond pixels using a learned similarity metric[EB]. 2015: 1512.09300.

[13] ABDULATIF S, CAO R Z, YANG B. CMGAN: conformer-based metric-GAN for monaural speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024 (32): 2477-2493.

[14] ZADOROZHNYY V, YE Q, KOISHIDA K. SCP-GAN: self-correcting discriminator optimization for training consistency preserving metric GAN on speech enhancement tasks[EB]. 2022: 2210.14474.

[15] ISOLA P, ZHU J Y, ZHOU T H, et al. Image-to-image translation with conditional adversarial networks[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Piscataway: IEEE Press, 2017: 5967-5976.

[16] YU G C, LI A D, WANG H, et al. DBT-net: dual-branch federative magnitude and phase estimation with attention-in-attention transformer for monaural speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022(30): 2629-2644.

[17] LU Y X, AI Y, LING Z H. MP-SENET: a speech enhancement model with parallel denoising of magnitude and phase spectra [J]. arXiv preprint arXiv: 2305.13686, 2023.

[18] VEAUX C, YAMAGISHI J, KING S. The voice bank corpus: Design, collection and data analysis of a large regional accent speech database[C]//Proceedings of the 2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE). Piscataway: IEEE Press, 2013: 1-4.

[19] THIEMANN J, ITO N, VINCENT E. The diverse environments multi-channel acoustic noise database(DEMAND): a database of multichannel environmental noise recordings[C]//Proceedings of the Meetings on Acoustics. ASA, AIP Publishing, 2013: 19(1).

[20] WENINGER F, ERDOGAN H, WATANABE S, et al. Speech enhancement with LSTM recurrent neural networks and its ap-

plication to noise-robust ASR[M]. Cham: Springer International Publishing, 2015: 91-99.

[21] Rec. ITU-T. Perceptual evaluation of speech quality(PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs[J]. Rec. ITU-T P. 862, 2001.

[22] TAAL C H, HENDRIKS R C, HEUSDENS R, et al. An algorithm for intelligibility prediction of time–frequency weighted noisy speech[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(7): 2125-2136.

[23] HU Y, LOIZOU P C. Evaluation of objective quality measures for speech enhancement[C]//Proceedings of the IEEE Transactions on Audio, Speech, and Language Processing. Piscataway: IEEE Press, 2008: 229-238.

[作者简介]

王鸿韬（2000– ），男，宁波大学信息科学与工程学院硕士生，主要研究方向为语音增强。

陆志华（1983– ），男，宁波大学信息科学与工程学院副教授、硕士生导师，主要研究方向为信号处理、多运动目标的实时跟踪等。

叶庆卫（1970– ），男，宁波大学信息科学与工程学院教授、硕士生导师，主要研究方向为信号检测、最优化搜索等。

章联军（1980– ），男，宁波大学信息科学与工程学院实验师，主要研究方向为网络技术实验通信与管理。